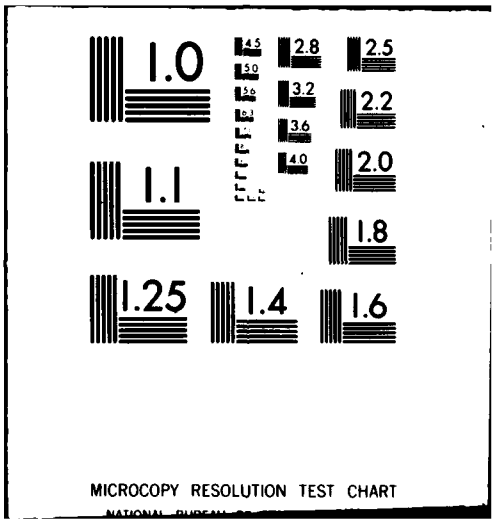
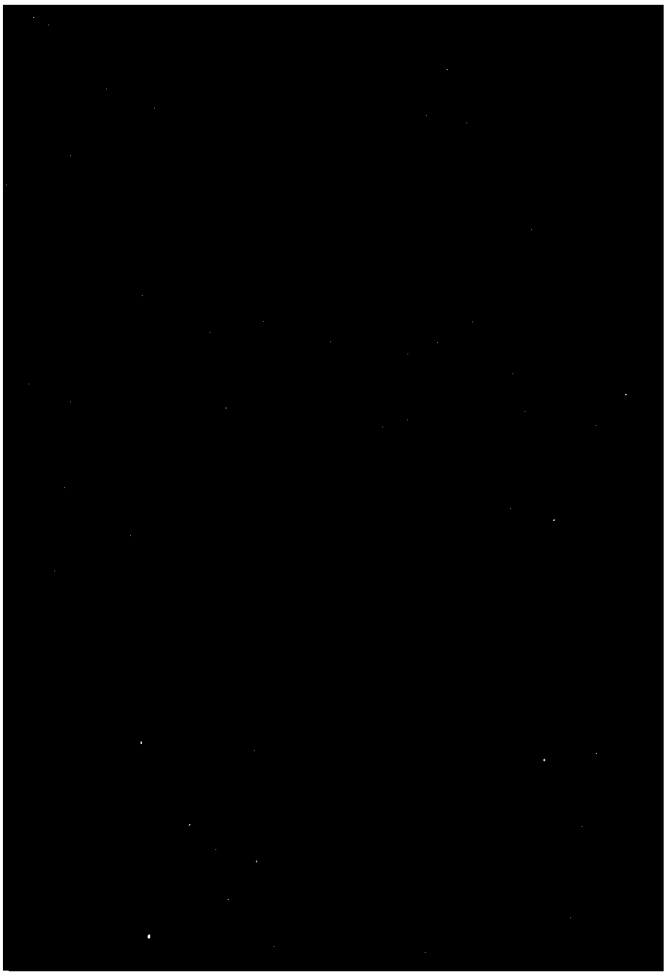
AD-A091 821 ROCHESTER UNIV N Y F/G 9/2
BIPLOT DISPLAY OF MULTIVARIATE MATRICES FOR INSPECTION OF DATA --ETC(U)
SEP 80 K R GABRIEL N00014-80-C-0387
UNCLASSIFIED TR-801 NL


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
MICROCOPY RESOLUTION TEST CHART



(6)

BIPLOT DISPLAY OF MULTIVARIATE MATRICES FOR INSPECTION OF DATA AND DIAGNOSIS

(9) Technical Report

BY

(14) TR-801

(10) K. Ruben Gabriel

Department of Statistics
and
Division of Biostatistics
Technical Report 801
University of Rochester
Rochester, New York 14642
USA

(12) 53

(11) Sept[illegible] 1980

Presented at the symposium, "Looking at Multivariate Data", at Sheffield, England, March 1980

(15) Supported by Contract N00014-80-C-0387 (new) from the Office of Naval Research. Reproduction in whole or in part is permitted for any purpose of the United States Government.

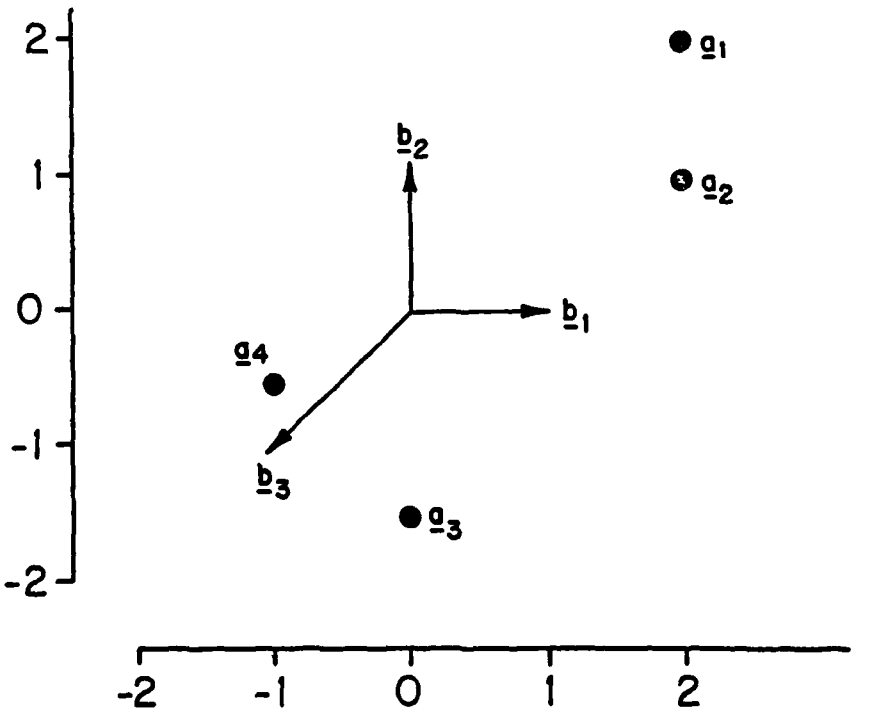
NSF → 307200

Display 1: A Biplot

Legend: ● $\underline{a}_u$ is u-th row marker
↗ $\underline{b}_v$ is v-th column marker

$$Y = A B'$$

$$\begin{bmatrix} 2 & 2 & -4 \\ 2 & 1 & -3 \\ 0 & -1\frac{1}{2} & 1\frac{1}{2} \\ -1 & -\frac{1}{2} & 1\frac{1}{2} \end{bmatrix} = \begin{bmatrix} 2 & 2 \\ 2 & 1 \\ 0 & -1\frac{1}{2} \\ -1 & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix}$$



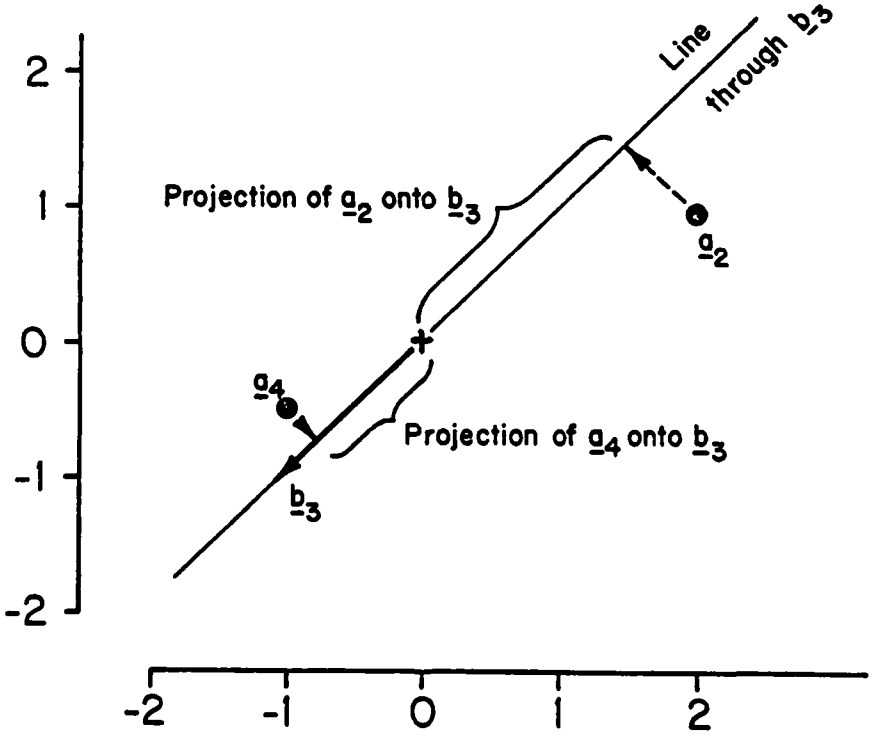
Display 2: Inner product representation of matrix elements on the biplot of Display 1

$$\left\{ \begin{array}{l} y_{2,3} = -(\text{Length of } \underline{b}_3) \times (\text{Length of projection of } \underline{a}_2 \text{ onto } \underline{b}_3) \\ y_{4,3} = (\text{Length of } \underline{b}_3) \times (\text{Length of projection of } \underline{a}_4 \text{ onto } \underline{b}_3) \end{array} \right\}$$

Third column of Y

$$\underline{y}_{(3)} = A \, \underline{b}_3$$

$$\begin{bmatrix} -4 \\ -3 \\ 1\frac{1}{2} \\ 1\frac{1}{2} \end{bmatrix} = \begin{bmatrix} 2 & 2 \\ 2 & 1 \\ 0 & -1\frac{1}{2} \\ -1 & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} -1 \\ -1 \end{bmatrix}$$



In this paper, I will discuss the biplot as a graphical multivariate technique, and I shall start by showing what a biplot is. I will then explain and illustrate its use in two applications: (1) in inspecting data matrices and (2) in diagnosing models to fit data. I will end by making some comments on advantages of this particular method as compared to other displays of multivariate data.

THE BIPLOT

A biplot (Gabriel, 1971, 1980) is a graphical display of a matrix Y of n rows and m columns by means of markers $\underline{a}_1, \underline{a}_2, \ldots, \underline{a}_n$ for its rows and markers $\underline{b}_1, \underline{b}_2, \ldots, \underline{b}_m$ for its columns. These markers are chosen in such a way that the inner product $\underline{a}_i'\underline{b}_j$ represents $y_{i,j}$, the i,j-th element of Y. Now, if we assemble all the $\underline{a}$ markers as rows of a matrix A and all the $\underline{b}$ markers as rows of a matrix B, then this inner product relationship means that matrix product AB' represents the matrix Y itself.

Let me make a remark about terminology. The prefix "bi" in "biplot" does not refer to its being two-dimensional but indicates that it is a joint display of rows and of columns of the matrix Y. When we have an analogous three-dimensional display, we refer to that as a "bimodel"; the prefix "bi" again indicates that it is a joint display of rows and columns; the ending "model" signifies that it is not plotted in the plane but uses further dimensions.

- Display 1 -

A simple example of a biplot is given in Display 1. The 4 by 3 matrix Y can be factorized as the product AB', A being

4 by 2, B' being 2 by 3. The biplot displays the rows of A, i.e., $\underline{a}_1', \underline{a}_2', \underline{a}_3'$ and $\underline{a}_4'$, as well as the rows of B, i.e., $\underline{b}_1', \underline{b}_2'$ and $\underline{b}_3'$. The first row of A, i.e., the vector (2,2) is displayed as the point $\underline{a}_1$; the second row (2,1) is displayed as the point $\underline{a}_2$, and the other two rows as points $\underline{a}_3$ and $\underline{a}_4$. The columns of B' are displayed as arrows $\underline{b}_1, \underline{b}_2$ and $\underline{b}_3$. The distinction between arrow display for columns and point display for rows is convenient: The viewer immediately sees which are row markers and which are column markers.

- Display 2 -

The inner product interpretation of this biplot can be seen from Display 2 which shows two of the elements of Y. Element $y_{2,3}$ is represented on the biplot by the inner product of $\underline{a}_2$ and $\underline{b}_3$. This inner product can be visualized by taking the direction through vector $\underline{b}_3$ and projecting the vector $\underline{a}_2$ onto it. The projection of $\underline{a}_2$ onto that direction is $3/\sqrt{2}$ units long; the length of $\underline{b}_3$ itself is $\sqrt{2}$ units long; the product is $3/\sqrt{2} \times \sqrt{2} = 3$; hence, the inner product is -3, the negative sign reflecting the projection's being in the direction opposite to that of the vector projected upon. Indeed, element $y_{2,3}$ is equal to -3. For another example, take element $y_{3,3}$: The inner product of $\underline{a}_3$ with $\underline{b}_3$ is visualized by projecting $\underline{a}_3$ onto the direction through $\underline{b}_3$. (This is the same direction that was used before.) The projection is of length $3/2\sqrt{2}$; the vector projected onto is of length $\sqrt{2}$; they are both in the same direction; therefore, the inner product is $+3/2\sqrt{2} \times \sqrt{2} = 1\ 1/2$, which is indeed the value of $y_{3,3}$.

The matrix Y could be biplotted exactly because it was of rank two. In general, an exact biplot of a matrix is possible only if the matrix is of rank one or two, because the biplot itself is planar. For a matrix of higher rank several steps have to be taken in order to display it by an approximate biplot. The first step is to approximate the matrix Y by a matrix $Y_{[2]}$ of rank 2. The second step is to factorize this rank 2 approximation $Y_{[2]}$ as a product AB' of a matrix $A_{(nx2)}$ and a matrix $B'_{(2xm)}$. The third step is to take each row of matrix A as a row marker a and each column of matrix B' as a column marker b. These markers are then plotted as an approximate biplot of the original matrix Y.

We next consider each of these three steps of approximation, factorization and display. The best known method for lower rank approximation is due to Householder and Young (1938). It minimizes the sum of squares of the deviations of elements of Y from elements of the reduced rank matrix $Y_{[2]}$. However, this method cannot be applied directly when weights are involved. The elegant mathematical relations that were used by Householder and Young break down as soon as one uses weighted least squares and multiplies the squared deviation $(y_{i,j}-y_{[2]i,j})^2$ by a weight $w_{i,j}$. An algorithm is available (Gabriel and Zamir, 1979), which allows this more general approximation. For a special kind of weights, Haber (1975) found an earlier solution. Another method of fitting lower rank matrices is by adaptive fits (McNeil and Tukey, 1975), and yet further methods might become available.

Factorization of the rank 2 approximation $Y_{[2]}$ is always

possible. Matrices $A_{(nx2)}$ and $B_{(mx2)}$ that satisfy $Y_{[2]} = AB'$ must exist. That follows from the definition of the rank of a matrix. However, such a factorization is not unique. In fact, if we post-multiply A by any 2 x 2 nonsingular R and pre-multiply B' by the inverse R^{-1}, the resulting $(AR)(R^{-1}B')$ factorization is just as valid as the original AB' factorization. We therefore have a choice as to which factorization to biplot. Note that transformation by a nonsingular matrix consists of a rotation of axes, a scaling along the new axes and another rotation, whereas the transformation by the inverse consists of the same rotations with a scaling which is reciprocal to the first one. This may help to give an idea of how different factorizations and different biplots are related. (An illustration of alternative factorizations and the resulting biplots was given by Gabriel, 1971).

The non-uniqueness of the factorization has some advantages for the statistician, who may choose a factorization which has desirable data analytic or statistical features. For instance, one particularly attractive factorization is referred to as the GH' factorization. This has orthonormal columns for G and therefore satisfies $Y'Y = HH'$, which is especially useful if the rows of Y represent individuals and the columns represent variables. Then $Y'Y$ is n times the estimated variance-covariance matrix, and so the inner products of the rows $\underline{h}$ of H in a GH' biplot represent the covariances, and the squared lengths of the $\underline{h}$'s represent the variances. The cosines between $\underline{h}$-vectors therefore represent the correlations between the variables. This biplot is useful in many statistical applications.

INSPECTION OF DATA

Next, I consider uses of the biplot. I will first describe the use of the biplot for inspecting data matrices. It is particularly useful for studying large data matrices, where eyeballing the large collection of numbers is quite impractical. Biplot display makes it much easier to see the main features of the matrix. I will illustrate this with a moderate size example because it is easier to present that in a paper. I should stress that I will not use the biplot to analyze the data statistically, and certainly not to test it for significance. Rather, I will use it for "looking at the data".

- Display 3 -

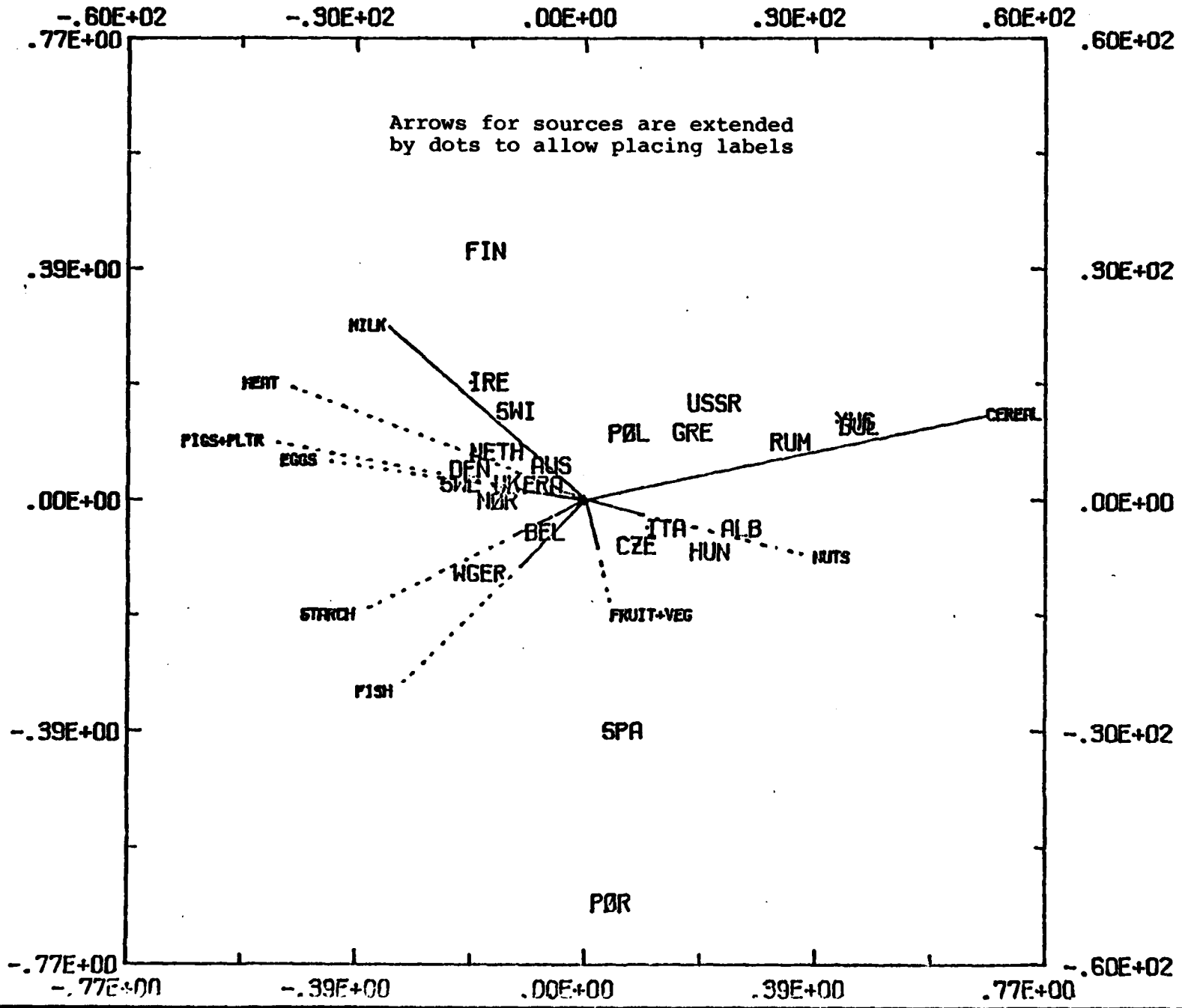
Display 3 shows the table of per capita protein consumption in 25 European countries: The rows are countries and the columns are nine different sources of protein. This matrix is biplotted in Display 4 after the mean of each column has been subtracted out. The points,or row markers, represent countries; the arrows, or column markers, represent sources of protein. This happens to be a GH'-biplot so that the lengths of the arrows represent the variances of the different sources of protein and the angles represent their correlations. The center of this biplot is at the European mean, or centroid, of all these sources of protein. The goodness of fit of $Y_{[2]}$ is of the order of 0.85; that is, the biplot displays 85% of the sum of squares of the mean centered data matrix Y.

- Display 4 -

Display 3: European protein consumption
(grams per head per day)

	Meat (Grazing animals)	Pigs and Poultry	Eggs	Milk	Fish	Cereals	Starchy Foods	Pulses Nuts, Oil-Seeds	Fruits Vegetables
Albania	10.10	1.40	0.50	8.90	0.20	42.30	0.60	5.50	1.70
Austria	8.90	14.00	4.30	19.90	2.10	28.00	3.60	1.30	4.30
Belg. Luxem.	13.50	9.30	4.10	17.50	4.50	26.60	5.70	2.10	4.00
Bulgaria	7.80	6.00	1.60	8.30	1.20	56.70	1.10	3.70	4.20
Czechoslovakia	9.70	11.40	2.80	12.50	2.00	34.30	5.00	1.10	4.00
Denmark	10.60	10.80	3.70	25.00	9.90	21.90	4.80	0.70	2.40
East Germany	8.40	11.60	3.70	11.10	5.40	24.60	6.50	0.80	3.60
Finland	9.50	4.90	2.70	33.70	5.80	26.30	5.10	1.00	1.40
France	18.00	9.90	3.30	19.50	5.70	28.10	4.80	2.40	6.50
Greece	10.20	3.00	2.80	17.60	5.90	41.70	2.20	7.80	6.50
Hungary	5.30	12.40	2.90	9.70	0.30	40.10	4.00	5.40	4.20
Ireland	13.90	10.00	4.70	25.80	2.20	24.00	6.20	1.60	2.90
Italy	9.00	5.10	2.90	13.70	3.40	36.80	2.10	4.30	6.70
Netherlands	9.50	13.60	3.60	23.40	2.50	22.40	4.20	1.80	3.70
Norway	9.40	4.70	2.70	23.30	9.70	23.00	4.60	1.60	2.70
Poland	6.90	10.20	2.70	19.30	3.00	36.10	5.90	2.00	6.60
Portugal	6.20	3.70	1.10	4.90	14.20	27.00	5.90	4.70	7.90
Rumania	6.20	6.30	1.50	11.10	1.00	49.60	3.10	5.30	2.80
Spain	7.10	3.40	3.10	8.60	7.00	29.20	5.70	5.90	7.20
Sweden	9.90	7.80	3.50	24.70	7.50	19.50	3.70	1.40	2.00
Switzerland	13.10	10.10	3.10	23.80	2.30	25.60	2.80	2.40	4.90
United Kingdom	17.40	5.70	4.70	20.60	4.30	24.30	4.70	3.40	3.30
USSR	9.30	4.60	2.10	16.60	3.00	43.60	6.40	3.40	2.90
West Germany	11.40	12.50	4.10	18.80	3.40	18.60	5.20	1.50	3.80
Yugoslavia	4.40	5.00	1.20	9.50	0.60	55.90	3.00	5.70	3.20
AVERAGE	9.83	7.90	2.94	17.11	4.28	32.25	4.28	3.07	4.14

Source: A. Weber (1973) Agrarpolitik im Spannungsfeld der internationalen Ernaehrungspolitik. Kiel, Institut fuer Agrarpolitik und Marktlehre (Mimeographed).


Display : GH'-biplot of European protein consumption
-.60E+02
-.30E+02
.00E+00
.30E+02
.60E+02
.77E+00
.39E+00
.00E+00
-.39E+00
-.77E+00
Arrows for sources are extended
by dots to allow placing labels
FIN
SPA
POR
MILK
MEAT
PIGS+PLTR
EGGS
STARCH
FISH
FRUIT+VEG
NUTS
CEREAL
IRE
SWI
USSR
POL
GRE
RUM
YUG
NETH
AUS
DEN
NOR
BEL
ITA
ALB
CZE
HUN
WGER

Looking at the configuration of the nine sources of protein, the most striking thing we see is that there is a very large variance for cereals and a somewhat large one for milk, but that the variances are relatively small for all the other sources of protein. The correlations are also interesting. On the left-hand side of the plot are all the animal sources of proteins; the angles between them are fairly small, which indicates high correlations between animal sources. Countries with high protein consumption from meat appear also to have high protein consumption from eggs, poultry, milk, etc. The marker for cereals is on the right side of the biplot, at an angle of about 180° to the markers for animal sources. Evidently, countries that have a high consumption of protein from animal sources have relatively low consumption of cereal protein and vice versa. Next, we note the markers for fruit and vegetables (and for fish (?)) to be at about 90° to both animal source and cereal markers. Apparently these sources of protein are pretty much uncorrelated with animal and cereal proteins.

It is interesting to consider which countries are typical of each source, i.e., which countries have high consumption of each kind of protein. For that purpose, the row markers can be displayed by means of three different symbols -- 1 for Western and Northern Europe, 2 for Eastern Europe, and 3 for Mediterranean countries. This simple device makes it easy to see that Eastern European countries are on the right of the biplot along with cereals; these countries consume much protein from cereals. Western and Northern European countries are on the left along with markers for animal protein. Mediterranean countries are partly towards the bottom of the biplot, which indicates that

fruit and vegetables, nuts, and fish are relatively important sources of protein for them.

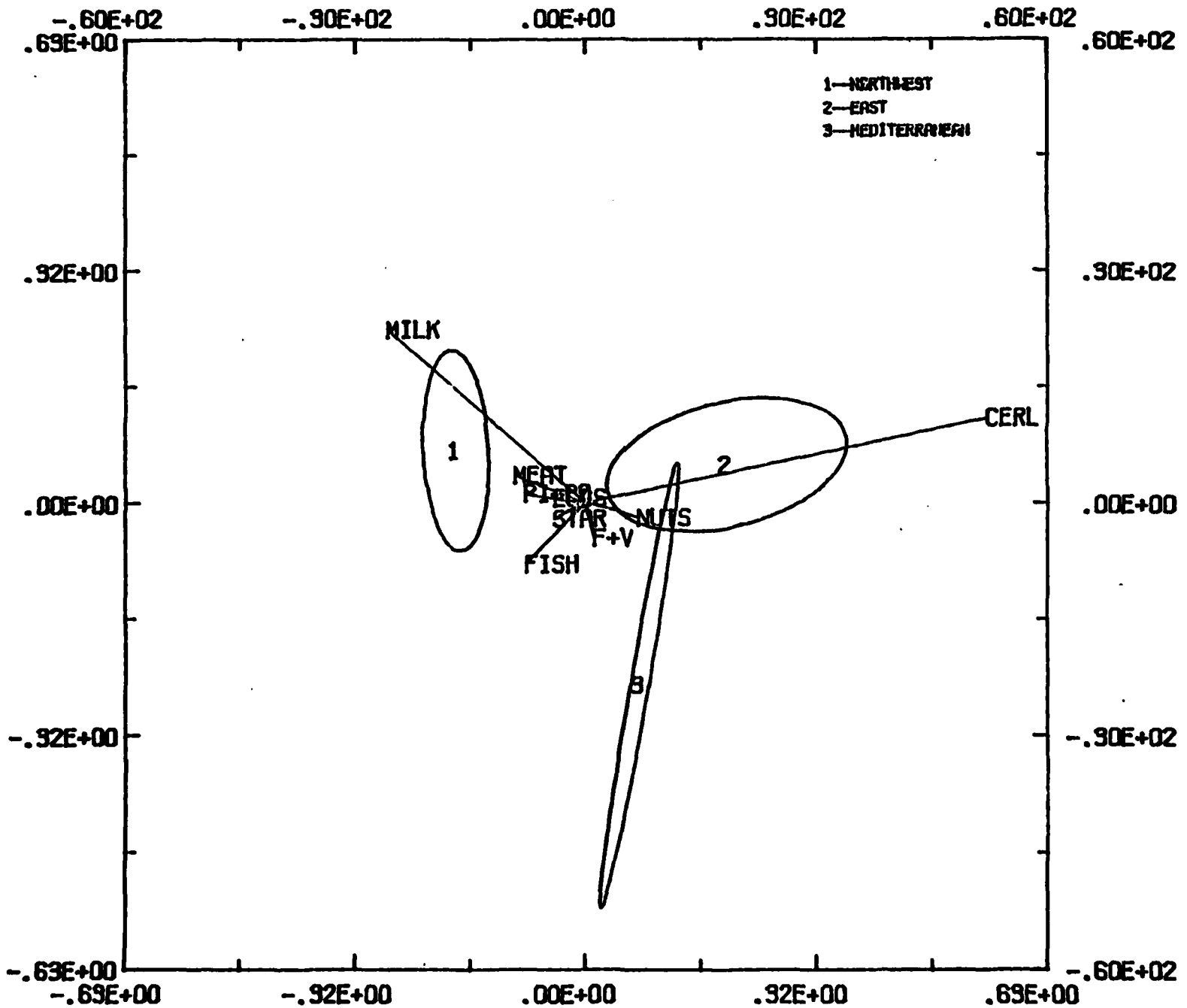
This example illustrates an important feature of the biplot. It displays not only the configuration of the variables, i.e., of the sources of protein, and the scatter of the individuals, i.e., of the countries, but it also relates the two. It therefore is able to reveal, for example, not only that consumption of cereal proteins is negatively correlated with consumption of animal proteins, but -- and this is the special feature of the biplot -- it also identifies countries which are typical users of cereal protein and countries which mostly use animal proteins. This joint display of countries and sources justifies the use of the prefix "bi".

Another method of displaying particular groups of countries on the biplot is the use of a concentration ellipse for the points of each group of interest. (A concentration ellipse is the two-dimensional analogue of a mean $\pm$SD interval: It is centered on the points' centroid and its "shadow" in any direction is a univariate mean $\pm$SD interval for the variate displayed in that direction; see Dempster, 1969, Ch. 7.) The usefulness of this concentration ellipse display is in summarizing a large number of points of each group by a simple figure.

- Display 5 -

The biplot of Display 4 is shown again in Display 5 with the countries' row markers replaced by concentration ellipses for the three groups. This very clearly shows the Northern and Western European group to be on the left, in the animal protein direction; the Eastern group on the right in the cereal direction

Display 5: GH'-biplot of European protein consumption with concentration ellipses for groups of countries



and the Mediterranean group to have a very elongated scatter in the nuts, fish, fruit and vegetable directions. It is obvious that the Eastern European group is much more heterogeneous than the Western and Northern European group and the shape of the Mediterranean scatter makes one doubt whether that should really be considered as a single group.

- Display 6 -

Use of concentration ellipses is of particular importance when large sets of data need to be displayed and there are more row markers than can be displayed effectively. Display 6 shows a biplot of breast tissue samples which were analyzed for enzyme activity. (The data are due to Dr. Russell Hilf of the Department of Biochemistry at the University of Rochester.) The activity of several enzymes and other phenomena were measured on each of several hundred breast tissue samples which had also been classified into four diagnostic groups: normal tissue (Group 4), cancerous tissue (Group 1) and two kinds of benign growths (Groups 2 and 3). When all the 700-odd points were displayed on the biplot, it was very difficult to distinguish the four groups of points. But the biplot with the concentration ellipses of the four groups -- Display 6 -- is much easier to grasp. One sees a clear distinction between the scatters of the cancerous and the normal tissues; each shows different enzyme activities. The two benign growth groups are intermediate between the preceding two in enzyme activities.

It is at times useful to consider only the variance-covariance configuration. Thus, in a GH' biplot one might omit

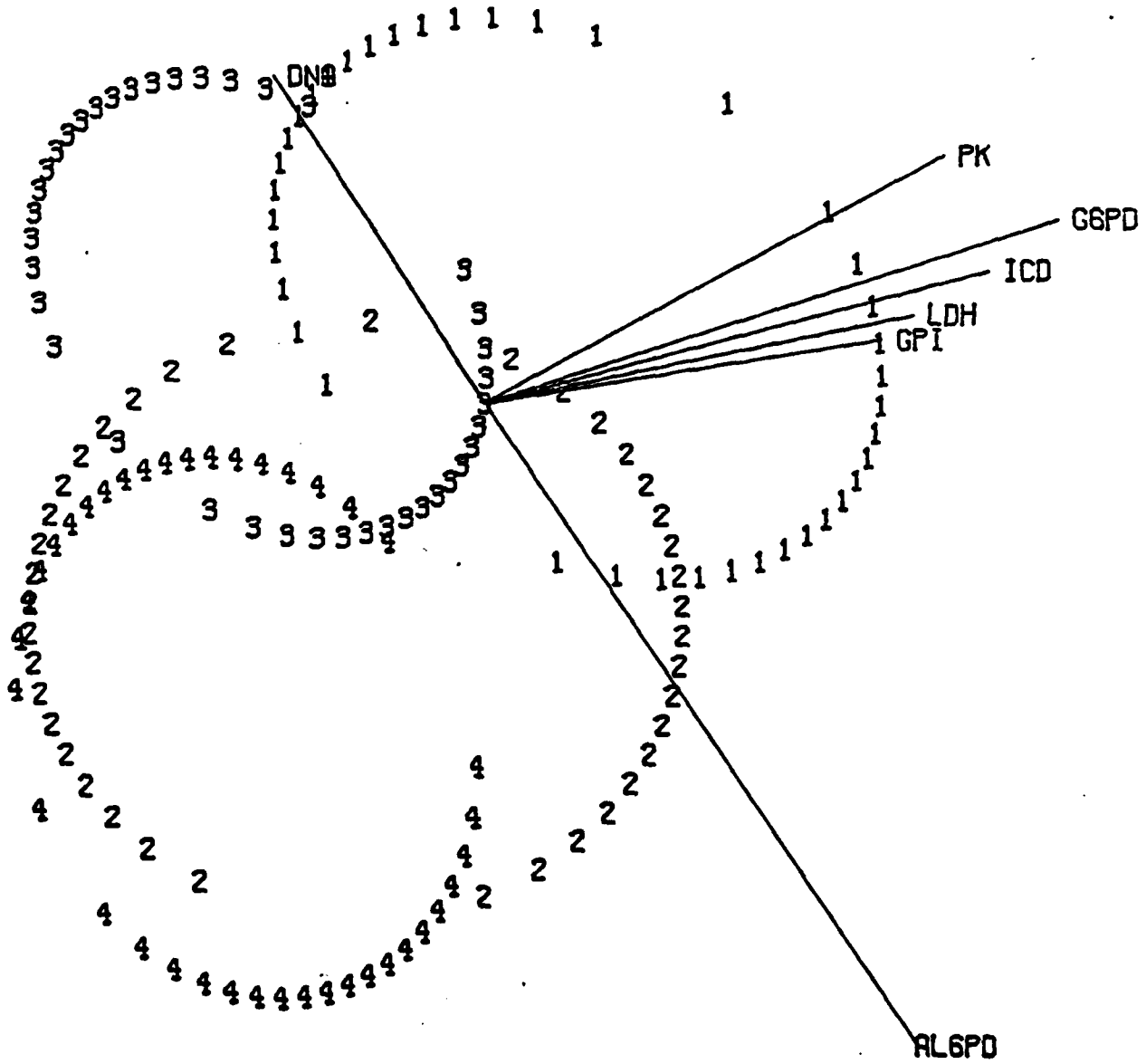
Display 6: Biplot of enzyme activity data for samples of breast tissue with concentration ellipses for four types of diagnosis

1--INFILTRATING DUCTAL CARCINØMA

2--FIBRØCYSTIC DISEASE

3--FIBRØADENØMA DISEASE

4--NØRMAL BREAST

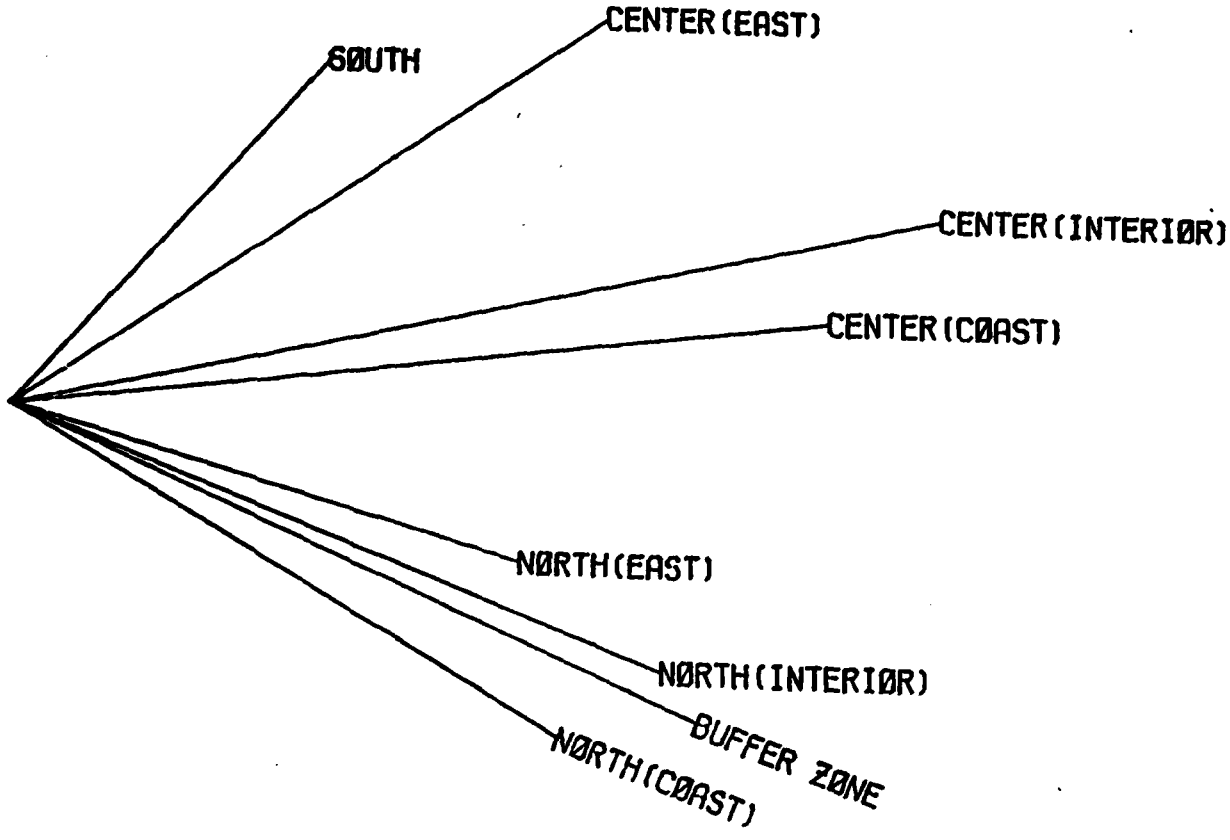


the markers for individuals (rows) and display only the variables (columns) h-markers. This will be referred to as a h-plot. One reason for wishing to ignore the individuals could be that they might be mere samples, or replicates, from a population -- and that it is only the population as an aggregate that is of interest. At times, one might want to use several h-plots and compare the variance-covariance configurations of several different populations.

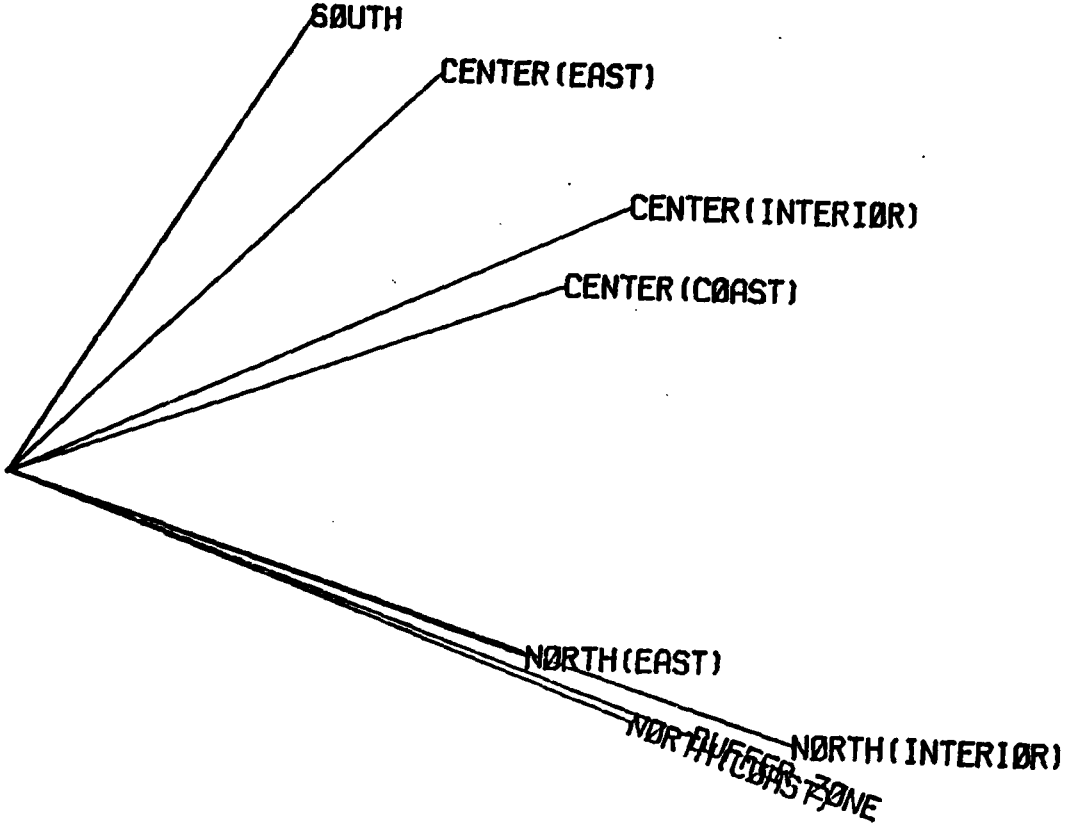
- Displays 7 & 8 -

An interesting example comes from the first randomized rainmaking experiment in Israel. Days were randomly allocated to have clouds seeded either in the North or in the Center of Israel. Displays 7 and 8 are h-plots of the precipitation in eight sub-areas of Israel -- Display 7 for Center-seeded days, Display 8 for North-seeded days (Corsten and Gabriel, 1976). The two h-configurations are, at first glance, very similar. At the top of each display is the h marker for the South, then come the markers for the three sub-areas of the Center of Israel, then, at the bottom of the displays, are the markers for the North of Israel, and for the "buffer zone" between the North and the Center. Both displays show that there was high correlation between sub-areas within the North, average correlation among the Center sub-areas and rather low correlation between the Center and the North.

Display 7: H-plot of variance-covariance of precipitation in Israel -- Center-seeded days



Display 8: H-plot of variance-covariance of precipitation in Israel -- North-seeded days



Display 9: Means and error sums of squares and products of anteater data

LOCALITY	NUMBER OF SKULLS	MEANS			SUBSPECIES
		z_1	z_2	z_3	
1. Sta. Marta, Columbia	21	2.054	2.066	1.621	Instabilis
2. Mina Geraes, Brazil	6	2.097	2.100	1.625	Chapadensis
3. Matto Grosso, Brazil	9	2.091	2.095	1.624	Chapadensis
4. Sta. Cruz. Bolivia	3	2.099	2.102	1.643	Chapadensis
5. Panama	4	2.092	2.110	1.703	Chiriquensis
6. Mexico	5	2.099	2.107	1.671	Mexicana
Total	48				

Within localities sum of squares and products (42 d.f.)

	z_1	z_2	z_3
z_1	0.013631	0.012769	0.016438
z_2	0.012769	0.012923	0.017135
z_3	0.016438	0.017135	0.036152

The variables z_1, z_2, z_3 are common logarithms of, respectively, basal length excluding the premaxilla, occipito nasal length and greatest length of nasals

Despite the overall similarity of the configurations of North-seeded and Center-seeded days, some differences are revealed by closer inspection of Displays 7 and 8. The most striking difference is that the correlations are considerably higher in the North when seeding was carried out in the North. Also, when one compares the lengths of vectors on the two h-plots, one readily sees that the variances of the Northern sub-areas were larger when the North was seeded whereas the variances of the Center sub-areas were larger when the Center was seeded. The explanation for these findings may be that the effect of seeding was (1) to make the seeded sub-areas more similar to each other, and (2) to augment the variance of rainfall in the seeded area (the means were also augmented -- though this is not shown on the h-plots).

- Display 9 -

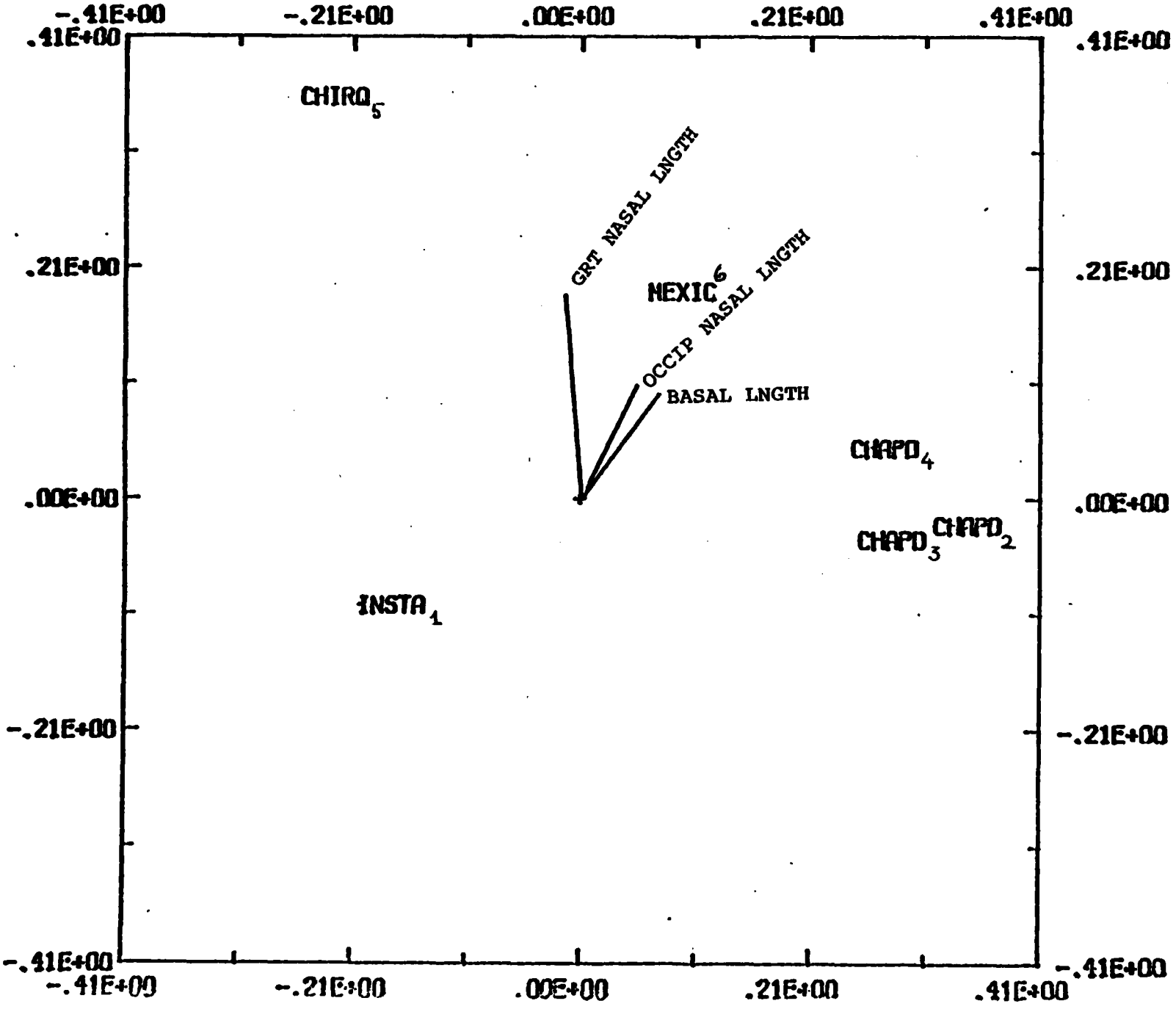
A somewhat more elaborate example is the data in Display 9 of three different cranial measurements of how subspecies of anteaters collected at six geographical locations (Reeve, 1940, quoted by Seal, 1964). The matrix that would be biplotted here is the six by three table of the six sample means of the logarithms of the three cranial measurements. Since these are averages of samples, it is appropriate when calculating their rank 2 approximation, to weight them by the sample sizes and the inverse of the within sum of squares and products matrix. This weighting is identical to that used in one-way multivariate analysis of variance (Gabriel, 1972).

- Display 10 -

On the resulting biplot, referred to as a JK'-biplot -- Display 10 -- each point represents a sample from one location and each arrow represents a log characteristic measured -- one of the three variables. What is immediately evident is that the three samples of sub-species Chapadensis are very similar -- they are very close together on this biplot. The location of Chiriquensis and the location of Instabilis are quite far from these three biplot locations and from each other. Mexicana is located between Chiriquensis and Chapadensis. Also, the general direction of the variables is up and slightly to the right, hence that is the direction of larger crania. This indicates that Instabilis is a smaller type of anteater, whereas Chiriquensis, Mexicana and Chapadensis are all larger. The difference between Chapadensis and Chiriquensis, on the other hand, is not one of overall size but one of a contrast between the different variables. Chiriquensis is relatively larger on the third variable -- greatest nasal length -- whereas Chapadensis is relatively larger on the first two variables. The two sub-species are thus seen to have different profiles of the variables.

This JK' biplot differs from the GH' biplots described above: Amongst other things, weights were used in fitting it. However, because of the particular weights used, biplot distances represent Mahalanobis distances between the different samples. Thus, the Mahalanobis distances between the Chapadensis samples are small; the one between Mexicana and Chapadensis is less then that between Chiriquensis and Chapadensis, etc.

Display 10: JK'-biplot of anteaters by subspecies and cranial measurements

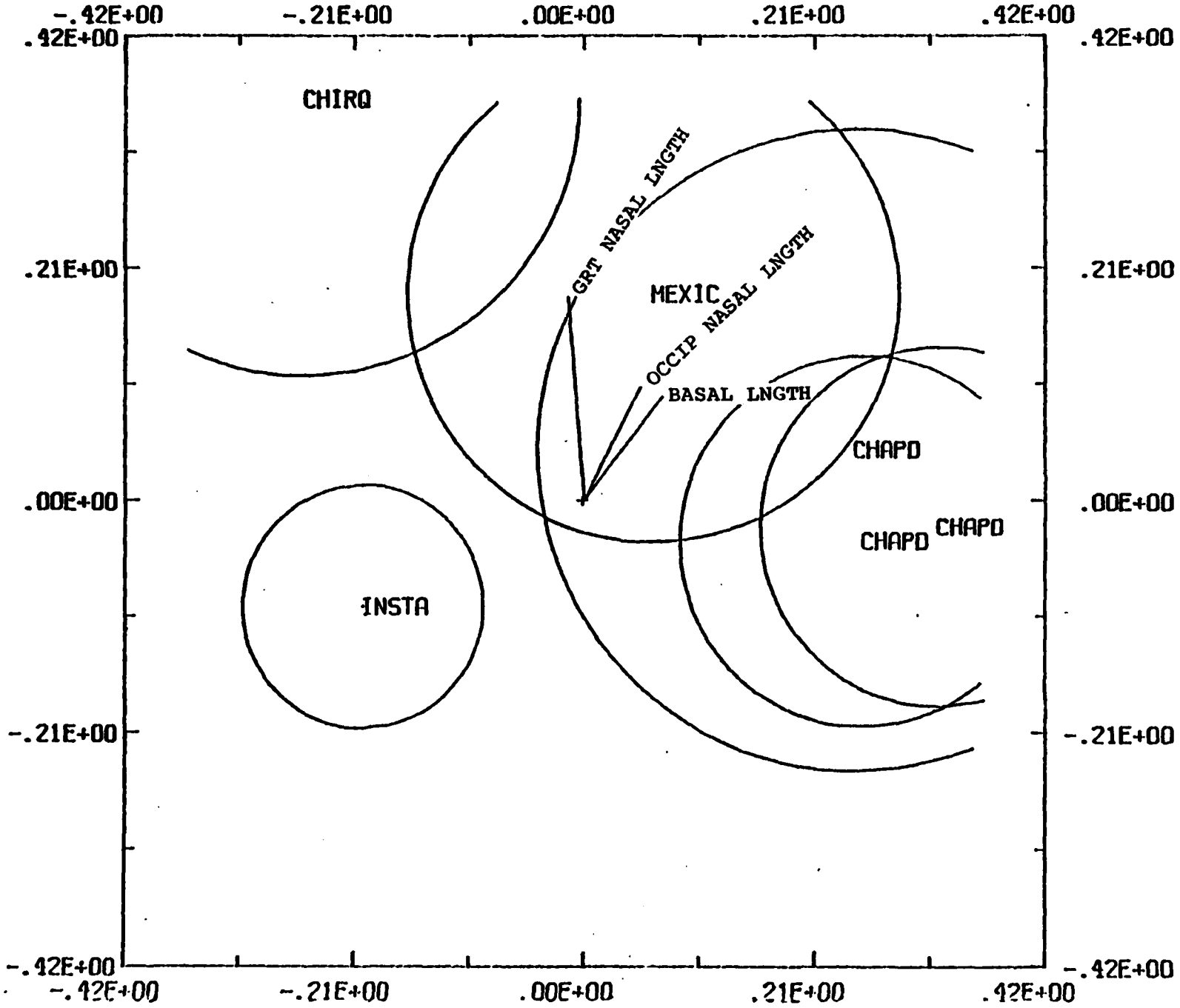


- Display 11 -

The Mahalanobis distance is closely related to Hotelling's T^2 except that the latter is scaled by the sum of the reciprocals of the sample sizes. It is possible to approximate the Hotelling T^2 test between pairs of samples by drawing circles around the biplot sample markers, where the radii of the circles depend on the critical point used for testing and on the sizes of the samples (Gabriel, 1972). This is illustrated in Display 11. The interpretation of these "comparison circles" is very obvious: Circles which intersect show a non-significant comparison; disjoint circles show a significant comparison. Thus, the three circles for Chapadensis overlap very much with each other and also with the Mexicana circle. This indicates that there are no significant differences between the three Chapadensis samples, nor between them and the Mexicana sample. Mexicana is not significantly different from Chiriquensis either. However, Mexicana is significantly different from Instabilis. In fact, Instabilis is found to be significantly different from everything else, and Chiriquensis is also different from Chapadensis. The general conclusions would be (1) that Instabilis indeed differs from the rest of the anteaters and is a smaller type; (2) the larger anteaters are of at least two groups: One containing Chiriquensis, the other Chapadensis; Mexicana could belong to either of these groups -- there are no significant differences which would indicate to which.

The graphical test used here is an approximation to Hotelling's T^2. In many cases such a Gaussian test may not

Display 11: JK'-biplot of anteaters with 5% experimentwise comparison circles



be valid and more robust tests might be needed. It might, for example, be possible to carry out re-randomization tests directly on biplots. We are currently trying out Mielke's (1976) "multiple response permutation procedures" for that purpose. Let me stress, however, that I see a very limited role for significance testing in the exploration of such multivariate data. In most multivariate situations, we have a fair number of samples and a fair number of variables; and we are rarely concerned with a test of an overall null hypothesis for all samples and all variables. Instead we usually want to find out what sort of differences exist and between which samples they occur. We are trying to explore rather than to test.

Multivariate analysis is essentially an exploratory technique rather than a confirmatory method. Indeed, by the time one gets to the stage of confirmation and sets up a well-defined null hypothesis for testing, one usually knows pretty well which particular variable, or what linear combination of variables,one is really interested in, so that the testing becomes univariate and not multivariate. I submit that multivariate analysis is principally exploratory and that techniques such as the biplot are usually very much more to the point than most tests of significance.

DIAGNOSIS OF MODELS

Another use of the biplot is that of diagnosing models which will fit a data matrix. This use is particularly important because statisticians really have very few techniques available for inspecting a data matrix and deciding what sort of model will fit it. Statistics textbooks have ample material on how to test a model once we have formulated it, but little or nothing on how to select a model, except by trial and error.

A biplot may be used to diagnose a model by looking for a pattern on the display and then infer mathematically what model that implies for the data matrix. For example, if the row markers are seen to be collinear, and the column markers are also noted to be collinear, and the two lines are at right angles to each other, one may infer that an additive model will fit the data closely, i.e., $y_{i,j} = \alpha_i + \beta_j$, for some set of alphas and betas. If, for another instance, one observes row markers and column markers to be on two non-perpendicular lines, one can infer that a concurrent model fits the data, i.e., $y_{ij} = \eta + \alpha_i\beta_j$, for some η, α_i's and β_j's. (This, by the way, is a reparametrization of Tukey's degree-of-freedom-for-non-additivity model.) Also, if one observes that all markers, for both rows and columns, are on one and the same line, it is obvious the matrix is a rank one and so the model is $y_{ij} = \alpha_i\beta_j$.

- Display 12 -

Display 12 shows these and some other rules of diagnosis derived by Bradu and Gabriel (1978). The first line indicates that when the row markers are collinear, the data may be fitted by a columns regression model. (This model is due to John Mandel (1961). It expresses each column as a linear regression on given row effects α_i). The next line of Display 12 similarly shows that, when the column markers are collinear, each row can be modelled as a linear regression on fixed column β's. When both row markers and column markers are collinear,

Display 12: Some biplot diagnostic rules

Row markers $\underline{a}_i$	Col. Markers $\underline{b}_j$	The model for $y_{i,j}$ is:
collinear	-	$\beta_j + \alpha_i \delta_j$ columns regression
-	collinear	$\alpha_i + \gamma_i \beta_j$ rows regression
collinear	collinear	$\mu + \gamma_i \delta_j$ concurrent (d.o.f.n.a.)
collinear	collinear	$\alpha_i + \beta_j$ additive
lines at 90° to each other		

(Bradu and Gabriel, 1978)

a concurrent model is diagnosed (as noted above), unless these two lines are at 90° to each other, in which case an additive model is diagnosed (as also noted above).

The rules of Display 12 apply even if some of the biplot markers are not on these lines. In such cases, the diagnoses apply to the subtable of the rows and columns whose markers are collinear. This is quite a remarkable feature of the biplot. It makes it possible to diagnose models not only for the entire matrix, but also for any sub-matrices. Most importantly, all these diagnostic indicators are very simple. The eye very easily picks up a straight line, even when it fits only some of the row markers or some of the column markers.

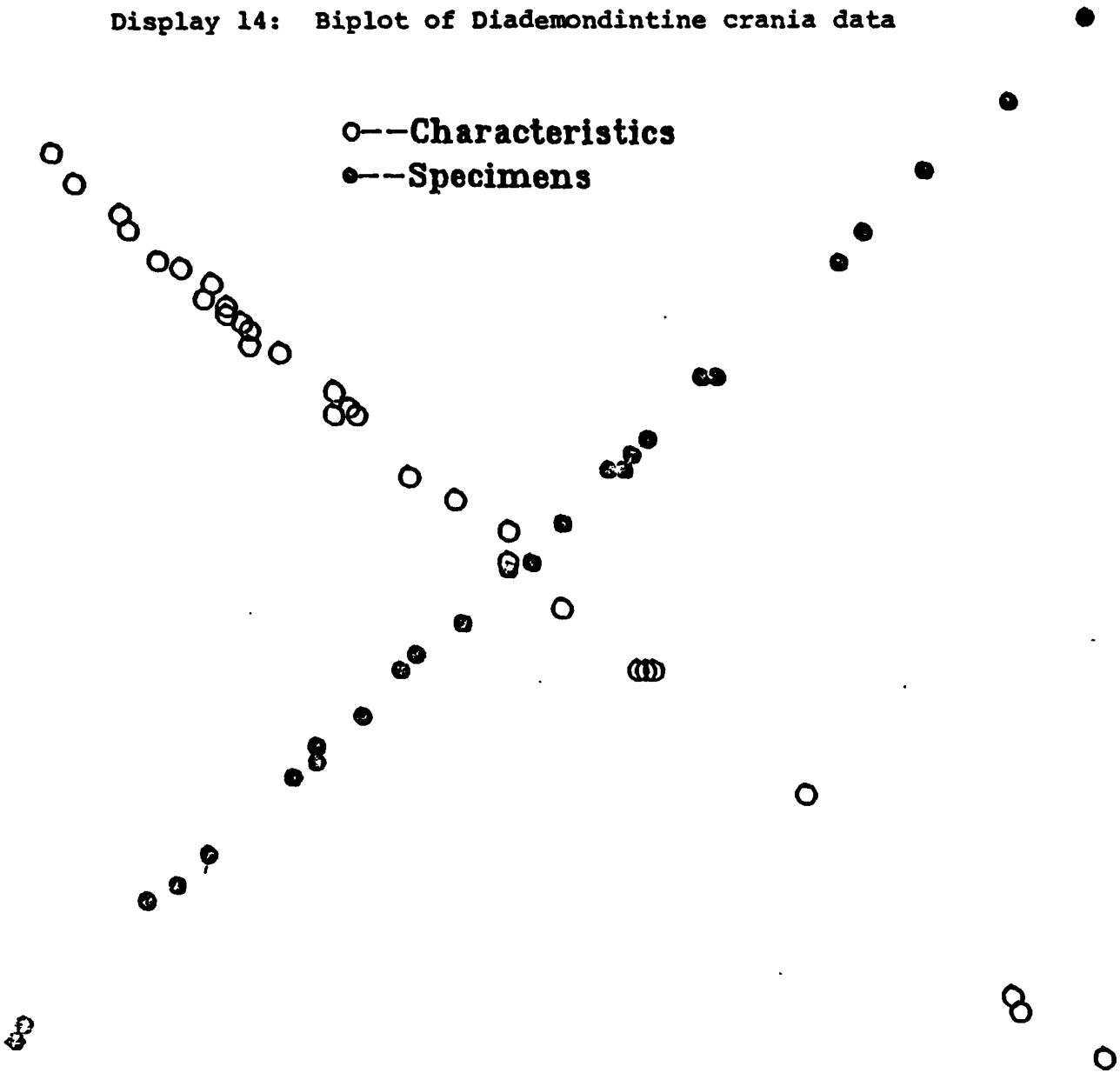
- Displays 13 & 14 -

Here are some further examples. Bradu and Grine (1979) considered cranial measurements for a number of specimens of fossils -- Display 13. This table has a large number of missing values, so that ordinary techniques for fitting were inappropriate. Bradu and Grine therefore used the algorithm developed by Gabriel and Zamir (1979) for weighted least squares and introduced 0 weights for the missing values and unit weights for present values. The resulting biplot is shown in Display 14. It is quite remarkable how closely the row markers cluster around one line and the column markers along another line. Using the diagnostics of Display 12, Bradu and Grine inferred that a concurrent model would fit the data very closely, as indeed it did. We note, however, that

Display 13: Thirty measurements on 26 specimens of Diademondintine crania

Variable	Specimen																									
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
1	275	230	287	177	---	---	134	74	92	112	142	---	---	84	47	127	193	205	260	232	130	172	51	247	197	187
2	305	230	331	194	---	---	147	76	99	120	154	---	170	---	53	135	---	238	280	253	139	185	---	258	209	209
3	130	98	141	83	73	76	55	39	45	48	58	---	74	36	23	61	90	95	122	90	58	80	19	98	89	82
4	158	118	171	108	---	---	86	36	52	61	88	63	---	44	21	70	103	110	143	142	72	92	32	149	104	105
5	177	131	216	108	---	---	93	44	57	69	95	---	97	---	33	78	---	143	163	163	81	105	---	160	122	125
6	138	---	145	92	97	77	64	38	52	65	71	---	79	---	29	72	93	104	130	114	69	85	27	115	99	100
7	96	---	97	62	67	61	42	24	33	43	49	---	52	---	19	47	63	69	84	76	47	60	19	74	65	71
8	107	88	119	---	75	67	54	31	---	42	61	---	---	---	---	56	82	88	103	93	55	68	---	---	85	75
9	114	78	131	64	---	---	54	22	29	37	60	48	60	---	17	41	---	---	64	95	42	52	---	---	54	57
10	160	138	187	125	108	---	83	43	53	69	90	---	116	58	32	92	126	---	167	144	86	100	32	145	123	---
11	260	205	303	184	---	---	136	70	82	109	141	---	167	---	50	133	---	223	263	241	131	163	---	---	200	194
18	253	144	207	130	---	---	115	66	65	93	115	92	106	---	28	63	---	---	230	218	100	138	---	---	174	153
19	200	126	197	105	128	---	96	61	60	76	93	68	92	48	26	52	---	---	166	160	74	115	---	180	---	---
20	59	42	58	37	39	32	29	24	22	25	28	27	---	24	10	24	41	---	56	53	26	36	11	53	39	37
21	48	42	46	31	37	29	28	18	20	---	24	---	32	17	9	---	33	42	47	40	---	30	11	36	34	---
22	64	56	68	37	52	41	30	22	22	26	29	---	36	---	9	---	41	49	68	58	---	37	12	48	43	---
23	23	21	---	18	17	---	13	12	8	12	---	---	13	10	6	---	14	---	25	17	12	13	8	17	15	---
24	64	47	72	35	51	---	36	17	18	28	40	33	31	20	6	17	---	---	70	78	31	43	---	---	49	41
26	159	98	---	66	---	---	75	44	54	58	77	56	---	---	12	34	---	---	132	111	62	87	19	142	---	99
28	93	---	99	54	80	58	45	28	39	44	48	44	48	34	16	32	64	64	92	77	45	55	18	90	67	65
29	33	29	36	20	30	19	19	12	16	---	15	16	18	12	8	---	27	22	36	22	17	23	8	30	23	23
30	58	49	62	35	42	39	30	19	21	25	28	---	33	---	9	---	37	24	60	54	26	35	10	56	38	37
37	83	55	---	52	51	---	41	32	---	24	32	30	---	---	16	37	---	---	---	75	---	---	---	---	---	---
38	62	52	82	53	42	38	32	28	---	25	29	30	37	34	15	31	45	---	69	63	---	46	---	---	46	---
39	53	50	71	47	35	32	32	23	---	24	28	24	30	22	13	31	42	53	66	49	---	36	14	54	43	42
42	30	34	44	27	---	26	24	14	20	23	28	23	---	16	13	27	33	---	36	40	25	28	---	40	33	---
43	25	32	35	31	---	20	20	15	14	18	21	20	---	16	8	22	25	---	32	24	14	21	---	31	30	---
47	17	12	20	8	11	12	9	3	4	6	---	---	8	3	3	6	11	14	---	18	---	10	---	22	---	10
51	100	80	120	71	---	---	44	28	23	28	39	37	41	---	13	36	---	80	86	69	34	55	13	---	69	---
52	69	54	62	35	---	44	38	---	---	38	34	---	29	---	---	27	49	50	73	63	34	42	---	56	50	50

Display 14: Biplot of Diademondintine crania data



the angle between the lines is very close to 90° and suspect that an additive model would also have fit pretty well.

- Display 15 -

A more complex example is data of Gamma radiation -- Display 15 -- classified by distance from the radiation source, number of intervening plates, the metal of which these plates consisted and two replications. This is a four-way layout so one has to confound several classifications in the rows and/or in the columns before one can display it in a biplot, as that is a matrix display. One way of doing this is to consider the data in the matrix form of Display 15, with the metals, distances and replications confounded in the rows, and only the number of plates appearing in the columns. This matrix -- after subtracting the overall mean -- is biplotted in Display 16, in which the column markers represent numbers of plates, and each row marker represents a combination of metal, distance and replication.

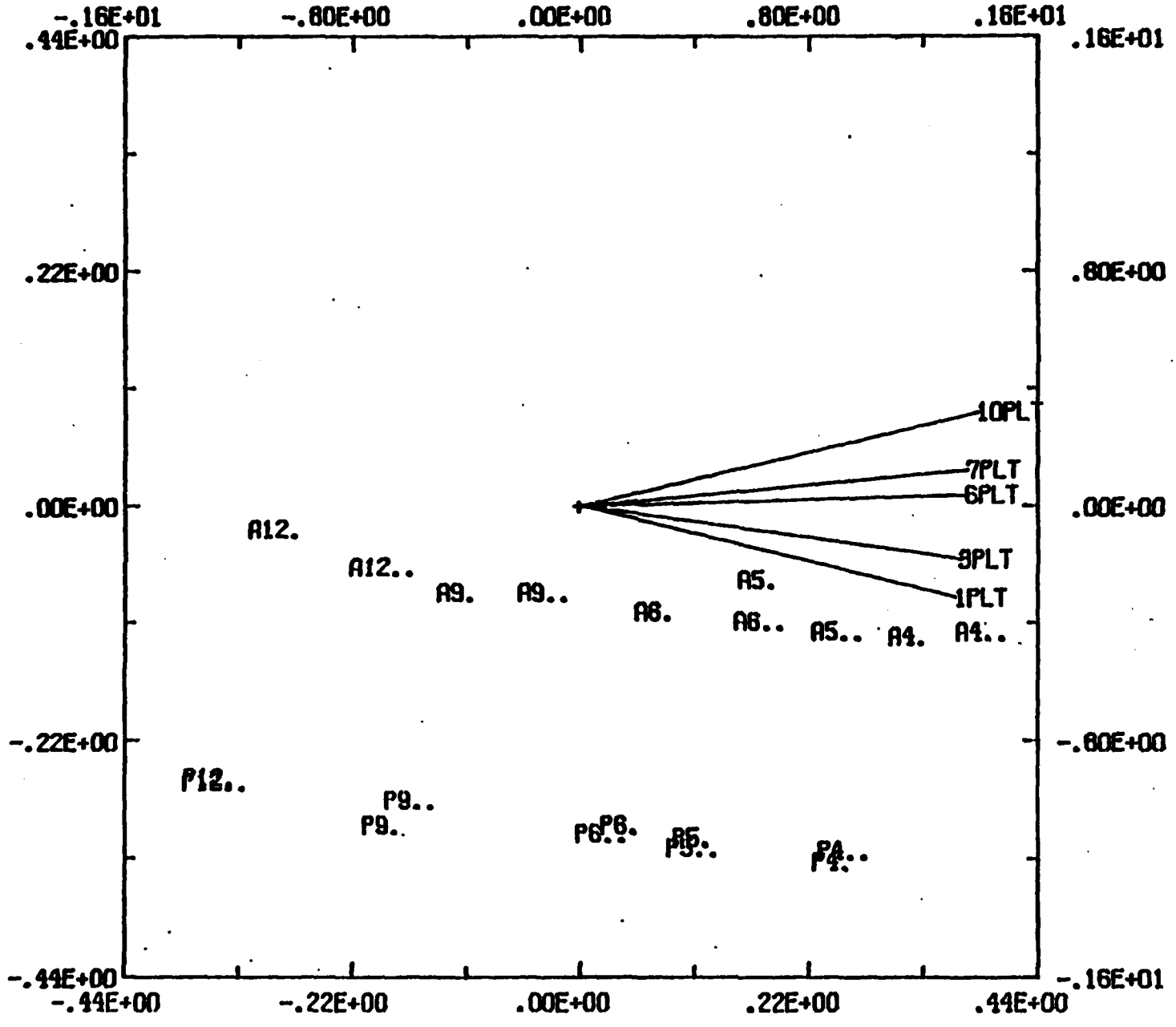
- Display 16 -

The biplot -- Display 16 -- of the radiation data clearly shows a linear pattern for the number of plates. The pattern for the row markers is not so immediately obvious. However, if for each of the ten distance x metal combinations, we average two replications, we find that these ten average markers lie on a non-rectangular lattice: The metals form two parallel lines and the distances form another five parallel lines. What model

Display 15: Absorption of gamma radiation by lead and aluminum

Row	Distance d in cm	Replic.	Number of plates p				
			1	3	6	7	10
					Lead		
1	3.8		1.801	1.765	1.696	1.670	1.606
2	5.2		1.621	1.572	1.516	1.486	1.425
3	6.0	I	1.526	1.481	1.406	1.401	1.333
4	9.0		1.222	1.169	1.102	1.078	1.010
5	12.5		0.973	0.939	0.862	0.850	0.781
6	3.8		1.805	1.768	1.704	1.680	1.615
7	5.2		1.609	1.572	1.511	1.482	1.408
8	6.0	II	1.494	1.461	1.387	1.324	1.315
9	9.0		1.233	1.208	1.130	1.111	1.046
10	12.5		0.978	0.930	0.870	0.844	0.779
					Aluminum		
11	3.8		1.834	1.818	1.811	1.790	1.777
12	5.2		1.632	1.613	1.600	1.603	1.597
13	6.0	I	1.509	1.482	1.476	1.454	1.447
14	9.0		1.249	1.224	1.204	1.211	1.179
15	12.5		0.976	0.971	0.966	0.960	0.943
16	3.8		1.916	1.913	1.884	1.887	1.871
17	5.2		1.732	1.723	1.698	1.696	1.674
18	6.0	II	1.632	1.624	1.592	1.588	1.579
19	9.0		1.344	1.341	1.312	1.311	1.290
20	12.5		1.118	1.118	1.106	1.086	1.066

Display 16: Biplot of gamma radiation data: Aluminum (A), Lead (P), Distance (number), number of plates (PLT), replications (.) or (..)



can be diagnosed from such a pattern? It may be useful to go through the algebraic steps of modelling for this case. Starting from any origin we can model the line for the column markers $\underline{b}_p = \underline{\alpha} + \lambda_p \underline{\beta}$, where λ_p is a parameter for the number of plates p and $\underline{\beta}$ is in the direction in which the column markers lie. We can also model the row markers for the average of the two replications as $\underline{a}_{m,d.} = \phi_d \underline{\gamma} + \psi_m \underline{\delta}$ with parameter ϕ_d depending on the distances d, and parameter ψ_m on the metals m. Vector $\underline{\gamma}$ would be in the direction of the parallels for the metals whereas $\underline{\delta}$ would be in the direction of the parallels for the distances. To see the form of the model for the data, we take the inner-product

$$\underline{a}'_{m,d.}\underline{b}_p = (\phi_d \underline{\gamma}' + \psi_m \underline{\delta}')(\underline{\alpha} + \lambda_p \underline{\beta})$$

$$= \phi_d \underline{\gamma}' \underline{\alpha} + \psi_m \underline{\delta}' \underline{\alpha} + \phi_d \lambda_p \underline{\gamma}' \underline{\beta} + \psi_m \lambda_p \underline{\delta}' \underline{\beta}.$$

This models the average $y_{d,m,p.}$ by an effect due to distance, plus an effect due to metal, plus two multiplicative effects, i.e., interaction terms, one of distance with plates and the other of metals with plates.

However, there is still more to be gleaned from the biplot of Display 16. The lines for lead and for aluminum are virtually parallel and pretty much at right angles to the line for plates. In terms of our parametrizations, this means the vector $\underline{\beta}$ is orthogonal to vector $\underline{\gamma}$. Therefore, the inner product $\underline{\beta}'\underline{\gamma}$ is zero and that term vanishes from the model. Defining $\pi_d = \phi_d \underline{\gamma}' \underline{\alpha}$, and $\sigma_p = \underline{\delta}' \underline{\alpha} + \lambda_p \underline{\delta}' \underline{\beta}$, one obtains the model $y_{d,m,p.} = \pi_d + \psi_m \sigma_p$. As ψ_m takes on only two values, this results in two additive submodels, one for each metal.

The distance effects π_d are the same for both metals, but the number of plate effects differ by a constant of proportionality. Indeed this kind of model could be fitted to these data. Note also that this example illustrates diagnosis for subtables: Rule four of Display 12 directly indicates an additive model for the data of each metal. (See Kester, 1979, for further rules.)

Further parametrization could be effected by noting that the distances along the lines through the column markers were pretty much proportional to the number of plates and therefore the parameters σ_p could be expressed as linear in the number of plates; similarly, the parallel lines for distances were spaced pretty much proportionally to the distances from the source of radiation and so π_d could be presented as a linear, or perhaps more precisely as a quadratic, expression in the distance d. The actual model that was fitted was an elaboration of the above and included linear and quadratic terms in the number of plates and in the distances from the source of radiation.

This is not only an instance of successful modelling but also shows the method by which a pattern observed on the biplot is translated algebraically into a model for the data.

All the models that we have diagnosed so far have been linear or bilinear in the various effects. It may be of interest to consider an instance in which such modelling was not sufficient. The example is one of mean monthly temperatures during the 24 months of 1951 and 1952 at 50 stations on the American continents (Brier and Meltesen, 1976) The data were biplotted -- Display 17 -- after the average temperatures for all 50 stations were subtracted out -- goodness-of-fit was 96%. (This analysis was carried out jointly with Mike Tsianco, 1980.)

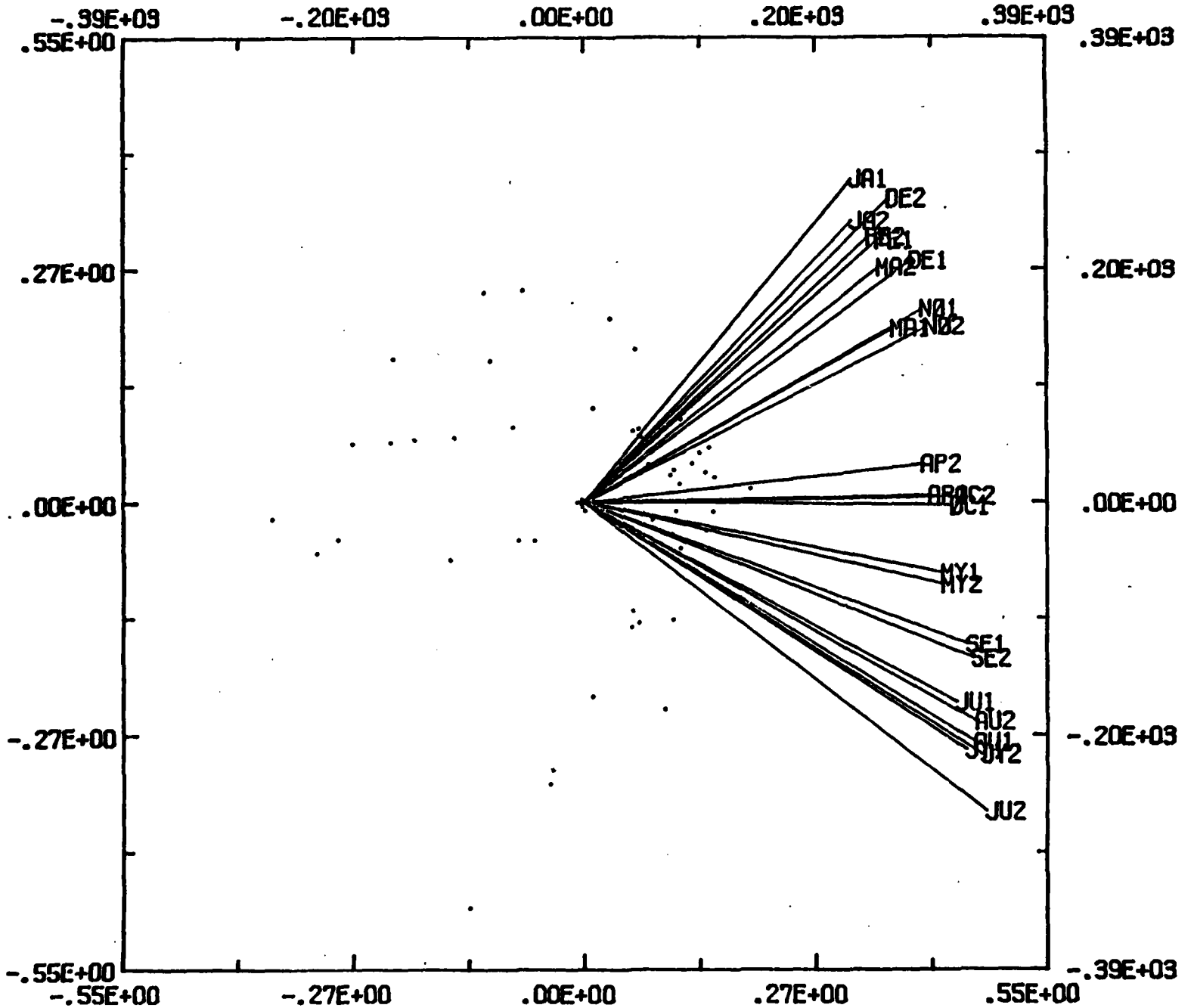
- Display 17 -

There is nothing particularly revealing about the scatter of row markers for the stations in Display 17. But the column markers fan out in a rather systematic manner: At first sight they would seem to be collinear and suggest a rows regression model. However, the order of the different months is interesting. It reveals a very similar configuration for the two years, with January at the top, then February and December, then March and November, and, somewhat farther down, April, October and May, then June, August, July and September. What sort of model does this suggest?

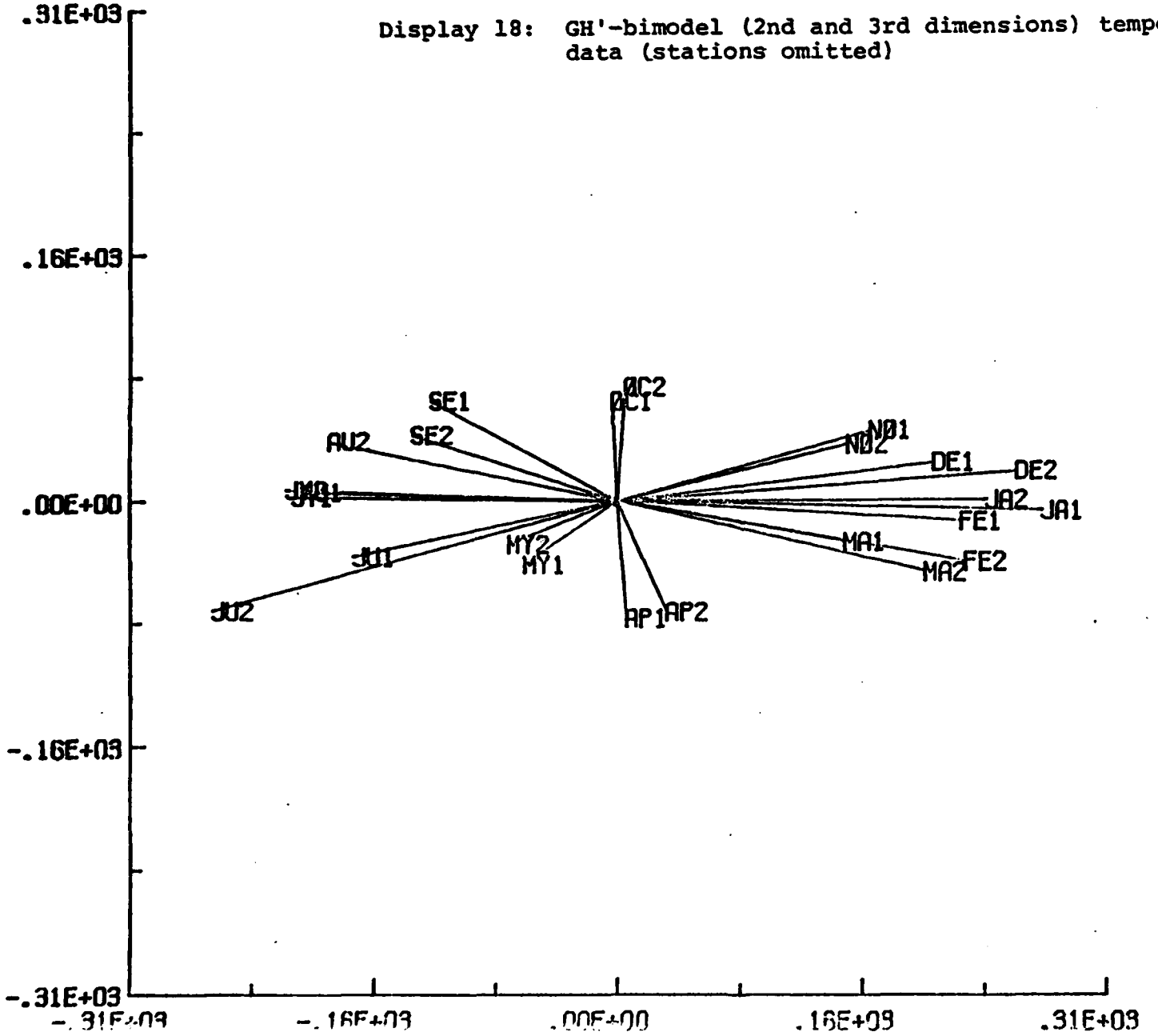
- Display 18 -

We note that the time sequence is systematic, going down from January to February, then to March and further down till June, then going up from July to December. This suggests that the time sequence may really be three-dimensional, the up-down-up movement on the biplot being complemented by a further change in a third dimension separating spring from fall. It is therefore worthwhile to fit a bimodel, i.e., a three-dimensional analogue of a biplot, and look at the plane of the second and third dimension, that is, essentially inspect the entire configuration from the right-hand side. This is shown in Display 18 in which the column markers are displayed on the plane of the second and third axes of the bimodel. Now we see a clear elliptical pattern from winter on the right, through spring at the bottom, summer on the left and fall on top.

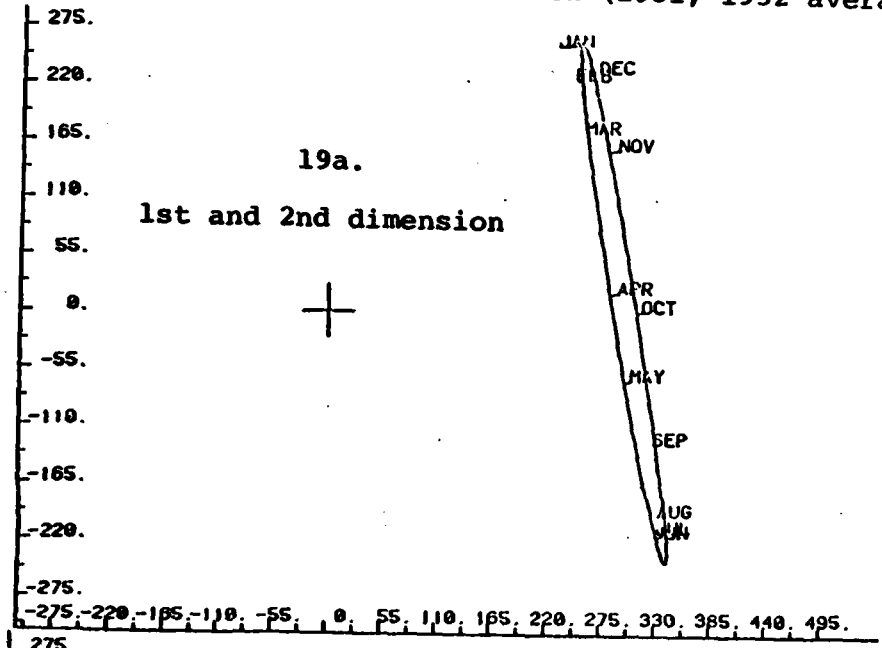
Display 17: GH'-bimodel (1st and 2nd dimensions) of temperature data: Dots - stations, lines - months



Display 18: GH'-bimodel (2nd and 3rd dimensions) temperature data (stations omitted)

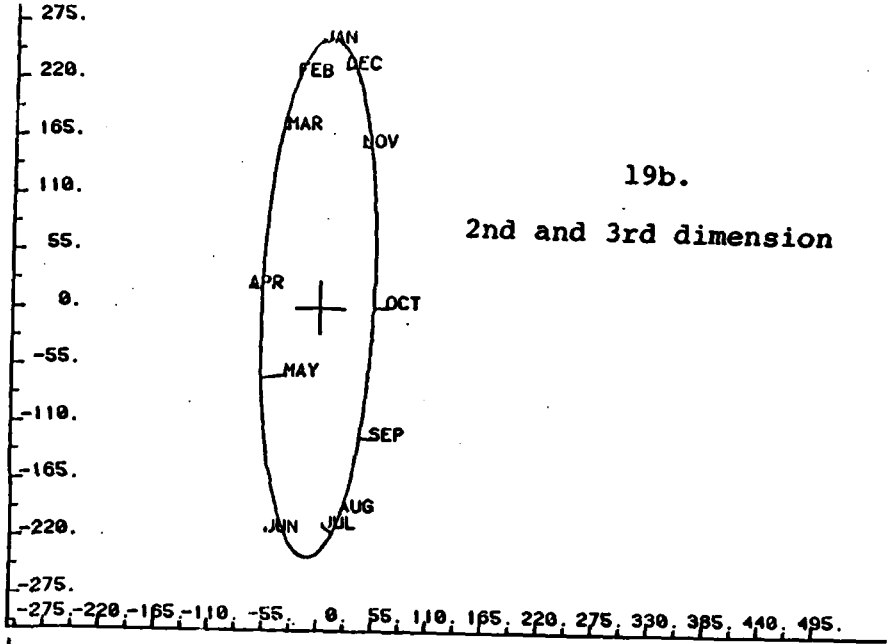


Display 19: Projections of ellipsoid fitted to monthly temperature data (1951, 1952 average)



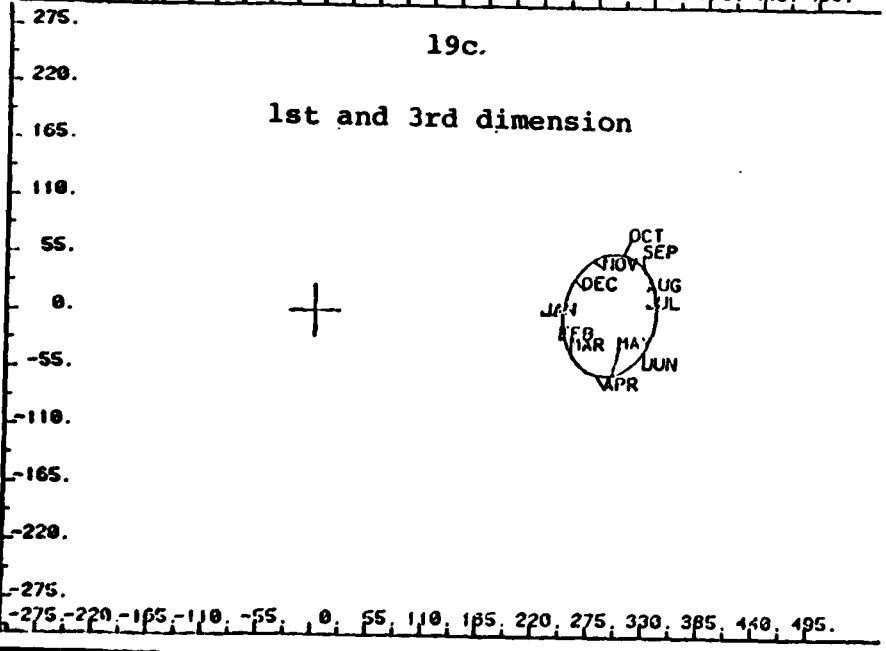
19a.

1st and 2nd dimension



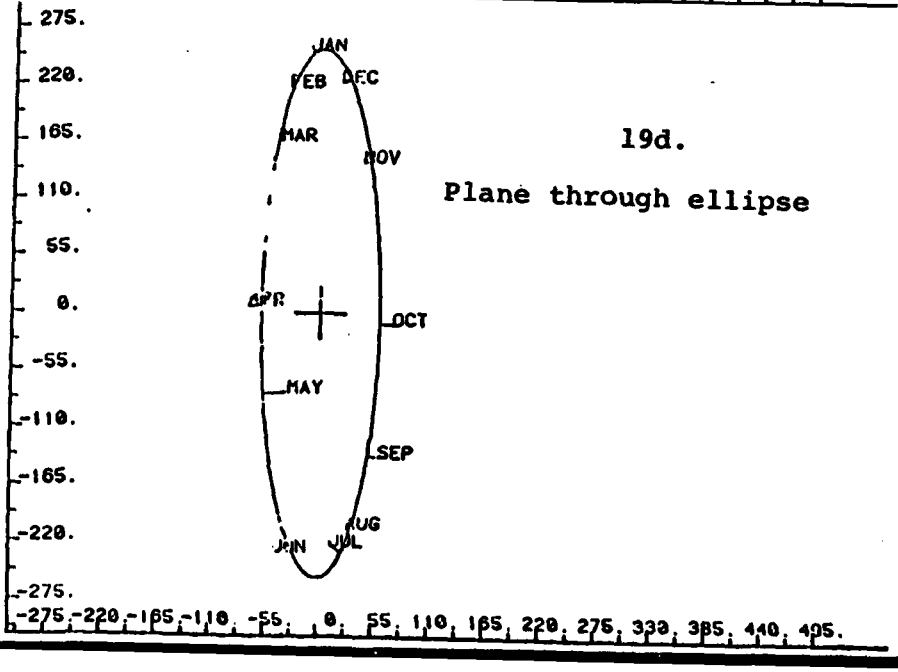
19b.

2nd and 3rd dimension



19c.

1st and 3rd dimension



19d.

Plane through ellipse

- Display 19 -

Another way of visualizing this is by means of ellipses fitted to the three-dimensional $\underline{h}$ configuration. Display 19 shows several projections of these fitted ellipses as viewed from the front, that is along the first and second dimensions; as viewed from the side, as we saw a moment ago; as viewed from the top; and as viewed orthogonally to the plane of the ellipse. (Note that in Display 19 the individual months are shown as averages for the two years and so there are only 12 markers instead of the 24 of Display 18. Also note that these are not concentration ellipses.)

What can be inferred from this elliptic pattern about models suitable for the data? What we are modelling by an ellipse is the configuration of the $\underline{h}$-vectors of the GH' bimodel. Y is displayed by rank 3 matrix product GH', and we are not considering the G factor but only the 24 columns of H, each having three elements. These $\underline{h}$'s have an elliptical configuration which we may model as follows, $\underline{h}_j = \underline{\mu} + \underline{\alpha}\cos\theta_j + \underline{\beta}\sin\theta_j$, for three-element vectors $\underline{\mu}$, $\underline{\alpha}$ and $\underline{\beta}$ and angles $\theta_1,\ldots,\theta_{24}$. This familiar parametrization of an ellipse represents the the center by $\underline{\mu}$, the major axis by $\underline{\alpha}$, the minor axis by $\underline{\beta}$ and the points along it by angles θ_j.

So far we have a model for the $\underline{h}$'s as observed on the bimodel. But our real concern is to obtain a model for the data matrix itself, i.e., for the elements $y_{i,j}$. Now, the bimodel representation is $y_{i,j} = \underline{g}_i'\underline{h}_j$ for a row $\underline{g}_i'$ of G and

one of the vectors $\underline{h}_j$. In view of the elliptical modelling of the latter, we obtain $y_{i,j} = \underline{g}_i'\underline{\mu} + \underline{g}_i'\underline{\alpha}\cos\theta_j + \underline{g}_i'\underline{\beta}\sin\theta_j$. This can be simplified by the following reparametrization:

$$\eta_i = \underline{g}_i'\underline{\mu},\ \psi_i\cos\phi_i = \underline{g}_i'\alpha \text{ and } \psi_i\sin\phi_i = \underline{g}_i'\underline{\beta}.$$

The model then becomes $y_{i,j} = \eta_i + \psi_i\cos\phi_i\cos\phi_j + \psi\sin\phi_i\sin\phi_j$, and, by the ordinary laws of trigonometry, that equals $y_{i,j} = \eta_i + \psi_i\cos(\phi_i + \theta_j)$. This simple harmonic model for the data has thus been shown to have been diagnosed by means of the bimodel.

This model makes a lot of meteorological sense. η_i is a station average temperature; ψ_i the amplitude of the annual harmonic variation in temperature, and when the model was fitted, these amplitudes were found to be larger farther from the equator and smaller close to the equator, as one would expect (Tsianco, 1980). The harmonic cosine element has its phases in terms of two arguments, ϕ_i depending on the station i and θ_j depending on the month j. Tsianco found θ_j to change from month to month by almost exactly $2\pi/12$, as one would expect from the annual cycle of temperature. He found the fitted values of ϕ_i to be much the same for all North American stations and again much the same for all South American stations -- the difference between the Northern and the Southern ϕ_i's was π -- which is what one would expect since it is well-known that it is warm in the North when it is cold in the South and vice versa.

This example has shown how inspection of a biplot/bimodel may lead to observation of a pattern which can be modelled and how such a model can lead to a model for the data themselves. It has also shown that the resulting model is in accord with

what we know about meteorology. Thus, biplot/bimodel inspection and consequent modelling for the data may give physically appropriate models.

SOME GENERAL COMMENTS

It may be in order to state the sequence in which I think display and modelling should be applied. One should begin by fitting the biplot or bimodel to a data matrix; from inspection of this display one might be able to infer a model or formulate a description of the data. Before one could conclude that this was an appropriate description, one should look at the residuals and ask whether they might be related to the fit of the biplot or the model, and/or whether they might be heteroscedastic. If so, one should look for forms of re-expression in the hope of yielding more homogeneous, less systematic, residuals from the next fit. This fitting, looking at residuals, re-expression sequence should be iterated until one is satisfied that the residuals are mainly noise.

Whilst doing these inspections of residuals, one should not merely look at general patterns of residuals but also spot outliers. In fact, this would seem to be an essential preliminary stage in all inspection of data. If there are extreme outliers, one must check the records from which the data came -- most of the time one would find gross errors which need to be corrected. In some cases, unexplained outliers would remain. It is extremely important to note unexplained outlying residuals in reporting analyses, even if they are omitted from the following fits and modelling because the methods of fitting might be

unduly influenced by them. Scientists who are interested in the data often find the outliers to be the most fascinating and instructive part of the whole data set. We, as applied mathematicians, enjoy finding patterns and fitting models and get the satisfaction of mathematical elegance of presentation of these regularities. But this may be of little interest to the scientists who are looking for new and unexpected phenomena rather than for neat formulation of patterns with which they are already familiar. It may well be that much of the progress of science is in finding the unexpected, the outliers, and being led to new ideas rather than in systematizing and parameterizing the familiar.

Let me make some final remarks about biplot display in comparison to a number of other techniques of data analysis. There are a number of steps in biplot display. (1) We start with a matrix Y. (2) We compute a reduced rank approximation $Y_{[2]}$; (3) We factorize that as $Y_{[2]} = AB'$; and then (4) we display the $\underline{a}$'s and $\underline{b}$'s in a biplot (or bimodel). Regularities that are in the original data can generally be expected to remain in the reduced rank approximation and therefore to be expressed in the factorization and to appear as patterns on the biplot. So matrix regularities will be displayed as biplot patterns. But scientific inference must proceed in the opposite direction. One observes the biplot patterns and tries to infer about the data. This is possible to the extent that the steps of approximation, factorization and display are reversible. Indeed,

display is reversed by visualization and factorization by inner-product multiplication. But the approximation step is reversible only as well as the goodness of fit of the reduced rank approximation. Often these approximations are very close and then one can say that the steps back from the biplot to the data can be retraced almost exactly. A number of examples have been presented above which show how one may parameterize a relationship, or pattern, on the biplot and then retrace the steps to see what model suits the data matrix.

The possibility of reproducing the data, at least approximately, from the biplot/bimodel display, is a unique feature of this particular method. There are a number of other methods, such as multidimensional scaling or correspondence analysis, in which one starts from a matrix, calculates a function of the matrix, e.g., interpoint distances, correlations, etc. and then produces some map of these distances or correlations by metric or non-metric methods. If there are regularities in the data then these maps of distances or correlations should reflect them. But we cannot even approximately retrace the step from the map of distances, or correlations, to the original data. This is because the distance, or correlation, functions which have been used to summarize the data are generally not one-to-one functions. Hence one cannot reproduce the data. One may model the distances, or correlations, but one cannot model the data by any of these other methods. The biplot seems to be unique in that it permits going back the extra step to the original data.

In summary, two main uses of the biplot have been presented. One is to inspect data matrices and look for patterns and relationships. In that use the biplot is very similar to several other methods. The other use of the biplot is to diagnose models to fit the data. For that purpose the biplot seems to be unique.

BIBLIOGRAPHY

Brier, G.W. and Meltesen, G.T. (1976). Eigenvector analysis for prediction of time series. Journal of Applied Meteorology, 15, 1307-1312.

Bradu, D. and Gabriel, K.R. (1978). The biplot as a diagnostic tool for models of two-way tables, Technometrics, 20, 47-68.

Bradu, D. and Grine, F.E. (1979). Multivariate Analysis of Diademontine Crania from South Africa and Zambia, South African Journal of Science, 75, 441-448.

Corsten, L.C.A. and Gabriel, K.R. (1976). Graphical exploration in comparing variance matrices, Biometrics, 32, 851-863.

Dempster, A.P. (1969). Continuous Multivariate Analysis, Reading, Mass.:Addison-Wesley.

Gabriel, K.R. (1971). The biplot - graphic display of matrices with application to principal component analysis, Biometrika, 58, 453-467.

Gabriel, K.R. (1972). Analysis of meteorological data by means of canonical decomposition and biplots, Journal of Applied Meteorology, 11, 1071-1077.

Gabriel, K.R. (1980). Biplot, Encyclopedia of Statistical Sciences, New York: Wiley.

Gabriel, K.R. and Zamir, S. (1979). Lower rank approximation of matrices by least squares with any choice of weights, Technometrics, 21, 489-498.

Haber, M. (1975). The singular value decomposition of random matrices, Ph.D. thesis at Hebrew University, Jerusalem.

Householder, A.S. and Young, G. (1938). Matrix approximation and latent roots, Am. Math. Monthly, 45, 165-171.

Kester, N. (1979). Diagnosing and fitting concurrent and related models for two-way and higher-way layouts, Ph.D. thesis at University of Rochester, New York.

Mandel, John (1961). Non-additivity in two-way analysis of variance, Journal Amer. Statist. Assoc., 56, 878-888.

Mandel, John (1969). The partitioning of interaction in analysis of variance, Journal Nat. Bur. Stand. (US), 73B, 309-328.

McNeil, D.R. and Tukey, J.W. (1975). Higher-order diagnosis of two-way tables, Biometrics, 31, 487-510.

Mielke, P.W., Berry, K.J. and Johnson, E.S. (1976). Multi-response permutation procedures for a priori classifications, Communications in Statistics. Theory-Methods, A5 (14), 1409-1424.

Reeve, E.C.R. (1940). Relative growth of anteaters, Proc. Zool. Soc. Lond., A110, 47-80.

Seal, H.L. (1964). Multivariate Statistical Analysis for Biologists, New York: Wiley.

Tsianco, M.C. (1980). Use of biplots and 3D-bimodels in diagnosing models for two-way tables. Ph.D. thesis at the University of Rochester, New York.

Mielke, P.W., Berry, K.J. and Johnson, E.S. (1976). Multi-response permutation procedures for a priori classifications, Communications in Statistics. Theory-Methods, A5 (14), 1409-1424.

Reeve, E.C.R. (1940). Relative growth of anteaters, Proc. Zool. Soc. Lond., A110, 47-80.

Seal, H.L. (1964). Multivariate Statistical Analysis for Biologists, New York: Wiley.

Tsianco, M.C. (1980). Use of biplots and 3D-bimodels in diagnosing models for two-way tables. Ph.D. thesis at the University of Rochester, New York.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: 801	2. GOVT ACCESSION NO.: AD-A092 821
4. TITLE (and Subtitle): The biplot for multivariate data analysis and diagnosis of models	5. TYPE OF REPORT & PERIOD COVERED: Technical Report
	6. PERFORMING ORG. REPORT NUMBER:
7. AUTHOR(s): K. Ruben Gabriel	8. CONTRACT OR GRANT NUMBER(s): N00014-80-C-0387
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Division of Biostatistics, University of Rochester Medical School, Rochester, NY 14642	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS:
11. CONTROLLING OFFICE NAME AND ADDRESS: Office of Naval Research, Arlington, Virginia 22217	12. REPORT DATE: September, 1980
	13. NUMBER OF PAGES: 46
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The biplot graphical display is introduced and its mathematical properties indicated. Its use for inspecting data matrices is illustrated for small and large matrices as well as for variance-covariance configurations and multivariate means of several samples. Its usefulness in diagnosing models is explained and illustrated by several examples requiring increasingly complex linear and harmonic models. (over)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-LF-014-6601

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. continued
Some comments are made on the relation to other multivariate displays.